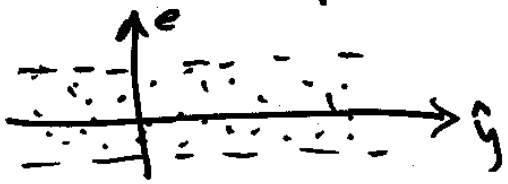
PRESS statistic = sum of squares of PRESS residuals

and $R^2_{PRESS} = 1 - \frac{\text{PRESS stat.}}{SST}$

(Compare to $R^2 = 1 - \frac{SSE}{SST}$)

Residual Diagnostics

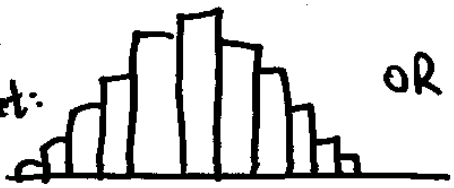
[1] Plot the residuals vs. the predicted values



Perfect:
- no change in dispersion
- no pattern

[2] Create a histogram or a box plot of the residuals

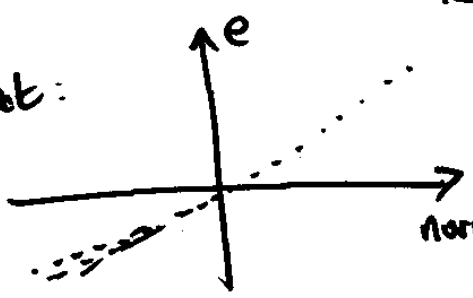
Perfect:

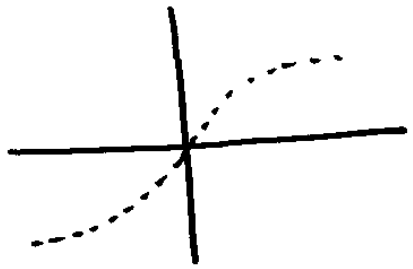


OR

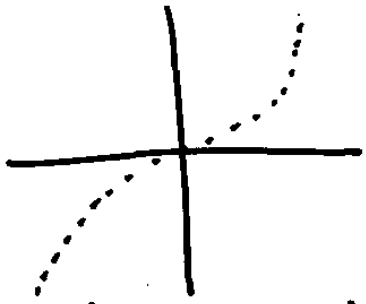
[3] Q-Q plot: Plot the residuals vs. the hypothesized normally distributed residuals

Perfect:



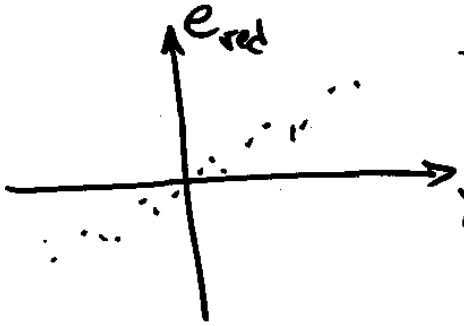


Your residuals have shorter tails than normal



Your residuals have longer tails than normal

[4] Remove 1 predictor from the model.
Run the reduced model & obtain the residuals.
Plot the residuals vs. the removed predictor.



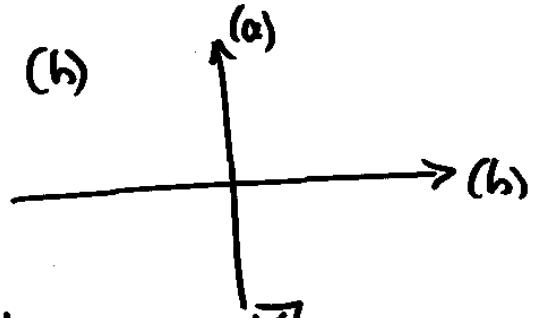
If you can see a pattern, then X_i should not have been removed from the model

If there is no discernible pattern, then
X_i is a candidate for removal from the model.

[5] Remove 1 predictor from the model.
Run the reduced model & obtain the residuals (a)
Regress the removed predictor on the remaining predictors & obtain those residuals (b)

(5)

Plot (a) vs (h)



Same interpretation as in [4]

[6] Keep track of order of data collection

$$\text{let } r_k = \frac{\sum_{t=k+1}^{n} e_t e_{t-k}}{\sum_{t=1}^{n} e_t^2} \quad \text{for } k = 1, 2, \ldots, n-1$$

(6)

These are the serial autocorrelations

Ideally, these are all 0.

Fact: Under the assumption of independent errors,

$$E[r_k] = 0 \quad \text{and} \quad V[r_k] \approx \frac{1}{n}$$

So Compute $\frac{r_k}{\sqrt{\frac{1}{n}}}$ & find any that are outside of ± 2

(7)

[7] Durbin-Watson test

$$D = \frac{\sum_{t=2}^{n} (e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}$$

(related to the $k=1$ case of [6])

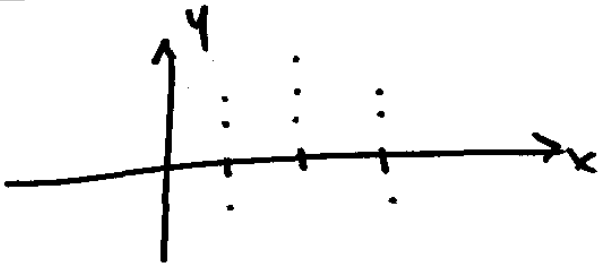
Under the assumption of independent errors,

$$E[D] = 2$$

If the DW test is significant,
you have found a violation of your assumption.

(8)

"Lack of fit" Test



Must have multiple observations of y
at at least 1 x-value

For simple linear regression, assume we have n observations, but m distinct x-values

x_1 : observe $y_{11}, \cdots, y_{1n_1}$ (9)

x_2 : $y_{21}, \cdots, y_{2n_2}$

$\vdots$

x_m : $y_{m1}, \cdots, y_{mn_m}$

$n_1 + n_2 + \cdots n_m = n$

$$SSE = \sum_{i=1}^{m} \sum_{j=1}^{n_i} (y_{ij} - \hat{y}_i)^2$$

$$= \sum_{i=1}^{m} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i + \bar{y}_i - \hat{y}_i)^2$$

$$= \sum_{i=1}^{m} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 + \sum_{i=1}^{m} n_i (\bar{y}_i - \hat{y}_i)^2 \quad (10)$$

(1) (2)

$$+ 2 \sum_{i=1}^{m} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)(\bar{y}_i - \hat{y}_i)$$

(3)

$$(3): 2 \sum_{i=1}^{m} \left[(\bar{y}_i - \hat{y}_i) \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i) \right]$$

$$\sum_{j=1}^{n_i} y_{ij} - n_i \bar{y}_i \quad (0)$$

(11)

$$\text{So } SSE = \underbrace{\sum_{i=1}^{m}\sum_{j=1}^{n_j}(y_{ij}-\bar{y}_i)^2}_{\text{"pure error"}\ SS_{PE}} + \underbrace{\sum_{i=1}^{m} n_j(\bar{y}_i-\hat{y}_i)^2}_{\text{"lack of fit"}\ SS_{LF}}$$

The lack of fit F test is

$$F = \frac{SS_{LF}/(m-2)}{SS_{PE}/(n-m)} \sim F_{m-2,\,n-m}$$

If this is significant, you have a poor fit.

(12)

Note about the df:

$(m-2)+(n-m) = n-2$, which was

df_E in simple linear regression

For multiple regression, $df_{LF} = m-p$

$df_{PE} = n-m$

So $(m-p)+(n-m) = n-p = df_E$